第 1 章　『現代日本語書き言葉均衡コーパス』入門

前川 喜久雄

1.1　はじめに

『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ）は、国立国語研究所が中心となって開発した日本語に関する初めての大規模均衡コーパスである。2011 年 8 月以来、BCCWJ は 2 種類の検索インターフェースを用いて、オンライン公開されている。全文検索専用のインターフェースは『少納言』（http://www.kotonoha.gr.jp/shonagon/ ）、形態素解析済データ検索用のインターフェースは『中納言』（https://chunagon.ninjal.ac.jp/）と呼ばれている。

これにくわえて、2011 年 12 月にはデータ全体を DVD に記録して公開した。これを以下では BCCWJ-DVD 版（Version 1.0）と呼ぶ。BCCWJ-DVD 版（Version 1.0）はその後広く内外で利用されたが、公開後早い時期から文境界の認定に問題があることが指摘されていた。また数字を桁単位に形態素解析するために導入した NumTrans（第 6 章参照）の仕組みについても、かえってデータの使い勝手を阻害しているとの指摘があった。

今般、これらの問題を中心にその他若干の問題を解消した新データを公開し、これを BCCWJ-DVD 版（Version 1.1）と呼ぶことにする。本文書は BCCWJ-DVD 版のマニュアルである。Version1.1 を公開するにあたり、本文書にも必要な改訂をくわえたので、タイトルを「『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版」に修正した。旧版（マニュアル第 1.0 版）と新版（同第 1.1 版）の主要な相違点は以下の 3 点である。

① 旧版では第 7 章を『中納言』の操作法にあてていたが、今回の改定に際して割愛した。『中納言』は毎年機能拡張を重ねて進化してきている。最新の操作法については『中納言』のオンラインマニュアルを参照していただきたい。
② 新版の第 8 章は新規に追加したもので、文境界の認定について BCCWJ-DVD 版（Version 1.0）から同（Version 1.1）への修正がどのように行われたかを説明している。
③ 旧版第 6 章では TSV データ（後述）と M-XML（Morphology-base XML）データ（後述）をまとめて解説したが、新版ではこれらを第 6 章（TSV）と第 9 章（M-XML）に分割した。

1.2　BCCWJ の特徴

1.2.1 均衡コーパス

BCCWJ は現代日本語の均衡コーパス（balanced corpus）である。現代日本語書き言葉

のできるだけ多くの変種をとりあげ、日本語の全体像を明らかにするための偏りのないサンプルを提供することを目標とした設計が施されている（第 2 章参照）。

BCCWJ は日本語に関する初の均衡コーパスであるが、その設計にあたっては、先行する諸外国の均衡コーパスを参考にしており、いくつかの点で先行コーパスに勝った設計がなされている。例えば、厳密な無作為抽出を可能なかぎり実施していること（第 3 章参照）、平均サンプル長を British National Corpus などに比べると短めに抑えることによって文献による語彙の偏りを低減していることなどである。

第 2 章および第 3 章で詳しく触れるが、BCCWJ は 3 個のサブコーパス、すなわち出版サブコーパス、図書館サブコーパス、特定目的サブコーパスから構成されている。

図 1-1 は、均衡コーパスが必要とされるひとつの事例を示している。この図は「食べ始める」「食べ続ける」のように用いられる補助動詞「～始める」「～続ける」が漢字を用いて表記される割合を BCCWJ のレジスター（register）（表 2-1 参照）ごとに示している。グラフ横軸に示されているレジスターについては 3.5 節以下参照。

最初に「～続ける」の結果を見ると、いずれのレジスターにおいても漢字表記率は 70%から 95%の水準にある。この場合、任意のレジスター、例えば新聞の分析によって得られた結論を他のレジスターに一般化することに大きな問題はない。

しかしながら「～始める」においては、レジスター間に顕著な差が存在している。そのため新聞データの分析から得られた結論は、雑誌・広報紙・教科書などのレジスターに対して一般化することができない。このような問題の存在は、均衡コーパスを分析することによって初めて知ることができるものである。

このようなレジスター間ないし語彙項目間の差は、あるいは何らかの一般的な要因に起因するものであり、従って予測可能であるかもしれない。しかし、そのような要因を発見するためにも均衡コーパスが必要とされるに違いない。

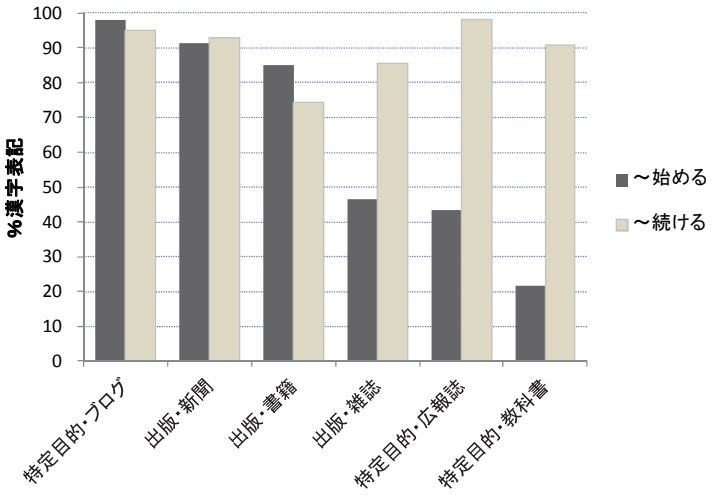

図 1-1: 補助動詞の漢字表記率のレジスター差（BCCWJ の解析結果）

もちろん BCCWJ にも種々の限界がある。例えば BCCWJ ではとりあげることのできなかった日本語書き言葉の重要なレジスターがある。その代表は漫画と広告である。これらのレジスターが現代日本語の動向（特にいわゆる新語の普及）に一定の影響を及ぼしていることは間違いない。しかし、画像情報への依存度が高いために他レジスターと同一の方法でのコーパス化が困難であること（この問題は雑誌サンプルの一部にも認められることがコーパスの構築過程で判明した）、および、著作権の処理に極度の困難が予想されることの二つの原因から、BCCWJ の収録対象とすることを断念した。

1.2.2 形態論情報

A. 短単位

BCCWJ にはアノテーションが施されている。最も重要なアノテーションは形態論情報、つまり文字列を語に分割して個々の語に品詞情報を付与した情報であろう。日本語のテキストは通常分かち書きされていないから、形態素解析されていないプレイン・テキストのデータから「国語」という文字列を単純に検索すると、目指す「国語」の他に「外国語」「韓国語」「中国語」「母国語」「自国語」等のごみが大量に生じてしまう。従来、日本語のコーパス言語学的分析では、正規表現を駆使して、プレイン・テキストから目指す文字列だけを得たり、後処理でごみを排除できることが研究者の基礎スキルとされてきたが、このような手法で常に目的を達することができるとは限らない。

正規表現を書くためにはあらかじめすべての表記上の可能性を把握しておかねばならないが、語によっては極端に異表記の多いものがある。例えば人名の「ヒロシ」には BCCWJ だけで 71 通りの表記がある。同じく、複合動詞の「ワキオコル」には終止形だけで 20 通り、活用形も含めると 324 通りの表記がある。これらの可能性をあらかじめ把握できる研究者は極めて稀であろう。

この問題は正しく形態素解析されたデータがあれば回避することができる。ただし日本語の形態論はいわゆる膠着語的な性格のために、「語」の規定に様々な問題が生じる。例えば「日本語」は全体でひとつの語とみてもよいが、「日本」と「語」の 2 語からなる複合語とみることもできる。

言うまでもないことだが、上記二つの解釈の間で言語分析上の絶対的な優劣を議論することには意味がない。重要なのは、どちらの解釈を採用するにしても、一旦ひとつの解釈を採用したら、その解釈の基礎となった言語学的観点を保持しながら、コーパス全体を分析できているかどうかである。

この点で従来の日本語形態素解析用辞書にはかなり深刻な問題が認められる。例えば ChaSen legacy の標準辞書として広く利用されている IPADIC では「国立国会図書館」は「国立＋国会図書館」の 2 語に分析されるが、「国立科学博物館」は「国立」「科学」「博物館」の 3 語に分析される。また「国立歴史民俗博物館」は「国立＋歴史民俗博物館」と 2

語に分析されるが、接尾辞「～学」を追加した「国立歴史民俗学博物館」は「国立＋歴史＋民俗＋学＋博物館」と 5 語に分析されてしまう。

言語学的な研究の基礎資料として用いるには、語認定におけるこのような一貫性の欠落は何としても回避したいところである。BCCWJ では上掲の例は以下のように分析される。（接）はその語の品詞が接尾辞であることを示しており、他はすべて名詞である。形態論的に一貫した言語単位が認定されていることがわかる。

国立国会図書館　　⇒　国立＋国会＋図書＋館（接）
国立科学博物館　　⇒　国立＋科学＋博物＋館（接）
国立歴史民俗博物館　⇒　国立＋歴史＋民俗＋博物＋館（接）
国立歴史民俗学博物館 ⇒　国立＋歴史＋民俗＋学（接）＋博物＋館（接）

BCCWJ が採用している形態論上のこの単位をわれわれは短単位（short unit word）と呼んでいる。短単位の認定基準については第 5 章参照。

B.　長単位（二重形態素解析）

短単位で一貫した言語単位を検索できるようになったのはよいが、短単位には副作用もある。上述の『中納言』を利用して、短単位の語彙素「ヒンディー語」を含む用例を検索すると 1 件もヒットしない。それならばと「中国語」を検索してみても同様である。これらの「語」は短単位としては「ヒンディー」と「語」、「中国」と「語」の 2 単位連鎖に分析されるからである。事実、これら 2 単位の共起関係を指定して検索すると、前者には 16 個、後者には 901 個の用例が見つかる。

しかし、これらの頻繁に利用される複合語を直接検索できないのは不便である。そこで BCCWJ には、主に複合語を把握する目的のために、長単位（long unit word）と呼ばれる単位に基づいた解析も施してある[1]。表 1-1 は同一のテキスト「公害紛争処理法における公害紛争処理の手続は」が短単位と長単位で、それぞれどのように解析されるかを比較したものである。

長単位の認定手順は第 5 章に詳しく説明されているのでここでは省略に従うが、結果として認定された長単位には以下のような特徴が認められる。

①複数の短単位から構成されている長単位には、「公害紛争処理法」のような実質語だけでなく、「における」のような機能語（いわゆる複合辞）がある。
②日本語のいわゆる膠着語的な性格を反映して「公害紛争処理」と「公害紛争処理法」がともに長単位として認定されている。BCCWJ を検索すると、さらに「公害紛争」「公害紛争処理制度」「公害紛争事件」「公害紛争処理機関」「公害紛争処理情報」等々

[1] 短単位と長単位による二重形態素解析は『日本語話し言葉コーパス』において最初に実施された。『日本語話し言葉コーパス』における短単位・長単位の定義と『現代日本語書き言葉均衡コーパス』における短単位・長単位の定義には、外来語の扱いなどに若干の相違があるが、大部分は一致している。

が長単位に認定されていることがわかる。

③長単位解析の結果は、短単位解析同様、解析対象テキストがもれなく長単位に分割されるという制約に従っている。そのため、いわゆる複合語（複合辞）だけが長単位に認定されるのではなく、短単位が単独で長単位に認定されることがある。表 1-1 の場合、最後の 3 行がこれに該当している。

表 1-1: 短単位と長単位の比較

短単位	短単位品詞	長単位	長単位品詞
公害	名詞-普通名詞-一般	公害紛争処理法	名詞-普通名詞-一般
紛争	名詞-普通名詞-サ変可能		
処理	名詞-普通名詞-サ変可能		
法	名詞-普通名詞-一般		
に	助詞-格助詞	における	助詞-格助詞
おけ	動詞-一般		
る	助動詞		
公害	名詞-普通名詞-一般	公害紛争処理	名詞-普通名詞-一般
紛争	名詞-普通名詞-サ変可能		
処理	名詞-普通名詞-サ変可能		
の	助詞-格助詞	の	助詞-格助詞
手続	名詞-普通名詞-サ変可能	手続	名詞-普通名詞-一般
は	助詞-係助詞	は	助詞-係助詞

短単位・長単位の認定基準を正確に理解するのは容易でないが、ユーザーは『中納言』の文字列検索機能を利用することで、検索したい文字列の単位構成についての知識を得ることができる。例えば「サーモンピンク色」が短単位としてどのように解析されるかを知りたければ、この文字列を文字列検索する際に、「結果表示単位」として「短単位」を指定すればよい[2]。検索結果の文字列には単位境界を示す縦線が挿入されて、以下のように表示される。

｜濃い｜サーモン｜ピンク｜色｜に｜なる｜。

また結果表示単位として「長単位」を指定した場合の表示は、

｜濃い｜サーモンピンク色｜に｜なる｜。

となるので、「サーモンピンク色」全体が 1 個の長単位として解析されていることがわかる。

C. 解析誤り

最後に、形態論情報について最も重要な情報に触れておく。形態論情報には解析誤りが含まれている。BCCWJ 全体の精度は 98%、コアデータ（第 2 章参照）に限れば 99%以上である。これは現在の形態素解析技術の最高水準を示す数字ではあるが、コアデータでも平均して 100 語に 1 語程度は誤りがあることになる。

解析誤りには、品詞を分類し間違えているもの、品詞は正解だが語彙素の細分類が誤っ

2 詳しくは『中納言』のオンラインマニュアル参照。

ているものなど、様々なタイプがある。もっとも深刻なのは、短単位境界そのものを分割し間違っている場合である。この場合、解析誤りが連続して出現することがあるので、注意が必要である。表 1-2 に解析誤りの例をいくつか示す。前文脈、後文脈中の縦線（｜）は短単位境界である。

表 1-2: 解析誤りの例

No	前文脈	キー	後文脈	語彙素読み	語彙素	品詞
(1)	｜ここ｜ん｜とこ｜、｜窮屈｜な｜こと｜ばかし｜で｜さ｜、｜	いやん	｜なっ｜ちゃう｜ったら｜ありゃ｜し｜ない｜...｜	イヤ	嫌	形状詞−一般
(2)	｜彼女｜は｜目｜を｜三角｜に｜し｜て｜部屋｜の｜中｜を｜	歩き	｜(まわっ)｜た｜。｜ルーク｜に｜この｜お｜礼｜は｜たっぷり｜し｜て｜あげる｜わ｜。	アルク	歩く	動詞−一般
(3)	｜奇妙｜な｜ほど｜	宮崎	｜(作品)｜に｜は｜家族｜、｜とりわけ｜親子｜関係｜の｜描写｜が｜避け｜られ｜て｜いる｜。	ミヤザキ	ミヤザキ	名詞−固有名詞−地名−一般

(1)は助動詞「に」の口語的な音便形を誤解析した例であり、短単位境界の認定誤りも生じている。(2)はいわゆる理論依存的な誤解析の例である。BCCWJ では「歩きまわる」全体が 1 個の短単位に分析されなければならないのだが（第 5 章参照）、このサンプルでは「まわる」が「歩く」から切り離されて 1 個の短単位に分析されている。(3)は短単位境界も語彙素の読みも正解だが、品詞分析で人名を地名に誤った例である。

誤解析の原因には様々なものがありうるが、BCCWJ の形態素解析では、コアデータを学習用コーパスとして解析器の機械学習を行っているので、学習用コーパスでカバーされていない語形の変異や品詞の細分類には対応が困難である。上例も学習用コーパスの限界による可能性が高い。

1.2.3 その他のアノテーション

形態論情報の他に、BCCWJ では文書構造と文字に関するアノテーションも提供されている（第 4 章参照）。これらは談話の研究や表記の研究に有益であると考えて施したアノテーションである。『中納言』では検索できないので、これらのアノテーションを利用するには BCCWJ-DVD 版が必要である。

また BCCWJ のサンプルには詳細な書誌情報が提供されている（第 7 章参照）。書誌情報はいわゆるメタ情報であり、言語の社会的側面の研究のために提供する情報である。書誌情報の一部は『中納言』の検索結果に表示されているが、『中納言』では書誌情報を検索条件に含めることはできない。書誌情報をキーとした検索を行うためには BCCWJ-DVD 版が必要である。

1.2.4 現代語

BCCWJ は現代語のコーパスであるが、ブラウンコーパスのように、或る特定の 1 年をきりとる形でデータを収集しているわけではない。一定の時間幅をもったサンプルが収録されており、その時間幅はサブコーパスないしレジスターによって変動している（表 3-1 参照）。

出版サブコーパスでは 2001 年から 2005 年までの 5 年間の幅であるが、図書館サブコーパスでは、これが 1986 年から 2005 年までの 20 年間に広がっている。特定目的サブコーパスに収められた種々のレジスター間にも相違があり、白書は 1976 年から 2005 年までの 30 年間をカバーしているのに対して、広報紙は 2008 年 1 年間だけである。すべてのレジスターが同一の時間幅をもっていることが望ましいのは言うまでもないが、実際にはデータの入手可能性が様々に異なることから、散らばりが生じている（第 2、3 章参照）。

1.2.5 著作権処理

コーパスの要件のひとつは、有償・無償を問わず、それが公開されていて誰でも利用できることである。そのためには、現代語コーパスの場合、著作権処理が必要になる。BCCWJ でもサンプルの性格に応じた著作権処理を実施した。

法律にはもともと著作権が存在しない。著作権が放棄されているテキスト（国会会議録と白書の一部）は、管理者にあたって著作権が放棄されていることを確認した。法人が著作権を有するテキスト（新聞記事、白書の大部分、雑誌記事の一部、広報紙等）は当該法人と交渉して許諾をもらった。

著作権の所属が明瞭でないテキスト（インターネット掲示板やブログ）の場合は、プロバイダ（Yahoo! Japan）の協力を得て、研究目的でデータを外部提供する可能性をネット上で告知した上で、告知の翌日以降に書き込まれたデータを提供してもらった。

個人の著作物のうち、権利者が日本文藝家協会等の著作権管理団体に所属しているものについては、管理団体の協力を得て、権利者に連絡をとることができた。しかし、例えば書籍の場合、このような方法で接触できる著者は全体の 2 割強であり、大部分のサンプルについては権利者の連絡先から調査を始める必要があった。

著作権データベース、各種紳士録、インターネット検索等で連絡先が判明することもあるが、それは例外的であり、多くの場合、連絡先を把握できない。その場合は、出版社に連絡をとって権利者への連絡を依頼するなどの方法で、多数の権利者と接触し、無償での利用を依頼した。

1.3 データの形式と内容

BCCWJ-DVD 版では、ユーザーの利便性に配慮して、サンプリングした言語データをさまざまな形式で提供している。Version 1.1 において提供されているデータは表 1-3 のとおりである。

NumTrans 非適用のデータは、第 6 章と第 9 章で説明するように Version 1.1 で新規追加されたデータである。最後の列（ディスク）に示したのは Version 1.1 を構成する 4 枚の DVD のうちどれにデータが保管されているかの情報である。さらに、この表には示していないが、書誌情報データとドキュメント類が Disc 1 に保存されている（1.5 節参照）。

表 1-3: BCCWJ-DVD 版（Version 1.1）に含まれる文書形式とデータの内容

文書形式	NumTrans	サンプル長	形態論情報	文書構造情報	ディスク
TSV	適用	統合	有	無†	Disc 2
	非適用	統合	有	無†	Disc 4
M-XML	適用	統合	有	有††	Disc 1
	非適用	統合	有	有††	Disc 3
C-XML	非適用	固定	無	有	Disc 1
	非適用	可変	無	有	Disc 1

† 文頭位置の情報（文頭ラベル）は提供されている（第 6 章参照）
†† C-XML（Character-base XML）の文書構造情報とは部分的に異なる（第 9 章参照）

(1) **TSV 形式と XML 形式**：データをタブ区切りテキストファイル（TSV）形式で提供しているか、タグ付き XML 文書として公開しているかの別である。TSV データは形態論情報を表形式で公開する目的に利用されており、短単位と長単位の情報は別のファイルに格納されている。XML 文書には 2 種類の別がある（下記(3) 参照）。

(2) **NumTrans 版と非 NumTrans 版**：「１９９９年」のように数字を含んだテキストを形態素解析するために事前に「千九百九十九年」のように形態素解析しやすい形にテキストを加工しているか（NumTrans 版）、していないか（非 NumTrans 版）の別である（第 9 章参照）。Version1.0 では NumTrans 版だけが公開されていたが、今回、非 NumTrans 版も追加公開する。NumTrans 版と非 NumTrans 版では、数字部分の短単位語数も形態論情報も異なることに注意が必要である。

(3) **C-XML 形式と M-XML 形式**：文書構造の情報だけを構造化したのが文書構造情報付き文字ベース XML（C-XML）である（第 4 章参照）。C-XML には後述する固定長（FIXED）サンプルと可変長（VARIABLE）サンプルの区別がある。形態論情報付き統合形式 XML（M-XML）は形態論情報を構造化したものであり、あわせて C-XML に含まれる文書構造情報の一部も構造化している（第 9 章参照）。

(4) **サンプル長**：BCCWJ のサンプルには固定長サンプル（1,000 字固定）と可変長サンプル（長さは様々。1 万字以内）がある。そしてレジスターによって、固定長と可変長の両サンプルを持つものと可変長サンプルだけのものとがある（第 2、3 章参照）。C-XML ではこれら両方のサンプルを別々に XML 化しているが（第 4 章）、一方、M-XML では、固定長と可変長を統合して重複部分を省いた統合形式サンプルに対して XML 化を

施している（第 9 章参照）。

(5) **コアデータと非コアデータ**：約 100 万短単位からなるコアデータに含まれるサンプル（コアサンプル）は、それ以外（非コアサンプル）に比べて形態論情報の解析精度が高い（第 5 章参照）。

(6) **書誌情報**：サンプルの書誌情報に関するメタデータを TSV 形式で提供している（第 7 章参照）。

(7) **文字符号化方式**:BCCWJ のすべての文書は文字符号化方式として UTF-8（BOM なし）を採用している。

図 1-2A-D に、BCCWJ-DVD 版（Version 1.1）の 4 枚のディスクのディレクトリ構成を示す。Disc 1（図 1-2A 参照）のルートディレクトリには 4 個のディレクトリがある。DOC ディレクトリ直下には、書誌情報データと著作権注釈情報データが格納されている。また DOC ディレクトリ下の MANUAL ディレクトリには、本文書、BCCWJ 構築時に蓄積したマニュアル類、さらに BCCWJ 公開後に出版された論文が格納されている。

書誌情報データについては第 7 章に詳しい説明がある。著作権注釈情報データは、権利者との交渉過程で、利用許諾に際して表示することを要請された注釈情報である。この情報は『中納言』でも当該サンプルがヒットした場合には表示される仕組みになっている。

C-XML には、文書構造タグ（第 4 章参照）を付したサンプルの XML データが、固定長（FIXED）と可変長（VARIABLE）に分かれて格納されている。

M-XML_NT （NumTrans）には、形態論情報付き統合形式 XML（M-XML）データ（第 9 章参照）が格納されている。この文書には固定長・可変長の区別はない。

C-XML 下の FIXED と VARIABLE および M-XML_NT の 3 ディレクトリの直下には、レジスターに対応するディレクトリがあり、各レジスターに属するサンプルが ZIP 圧縮されている（圧縮の方式については後述）。FIXED 直下のディレクトリは 書籍（PB）、雑誌（PM）、新聞（PN）、図書館 SC（LB）、白書（OW）の 5 個だけであるが、VARIABLE と M-XML_NT ディレクトリ直下には 13 個のディレクトリが存在する。CORE_NT ディレクトリについてはすぐ後で触れる。

Disc 2 は NumTrans 版の TSV データを格納している（図 1-2B 参照）。短単位（TSV_SUW_NT）、長単位（TSV_LUW_NT）の各ディレクトリ直下に、Disc 1 と同様に 13 のレジスターごとに圧縮されたデータが格納されている。

Disc 1 の CORE_NT ディレクトリには、BCCWJ コア（第 2 章参照）の対象となったサンプルの M-XML データ の NumTrans 版と TSV データの NumTrans 版（短単位と長単位）が格納されている。これはコアだけを処理したいユーザーの便宜を図ったものであり、このディレクトリのデータはすべて、Disc 1 の M-XML_NT、Disc 2 の TSV_SUW_NT、TSV_LUW_NT と重複して格納されている。

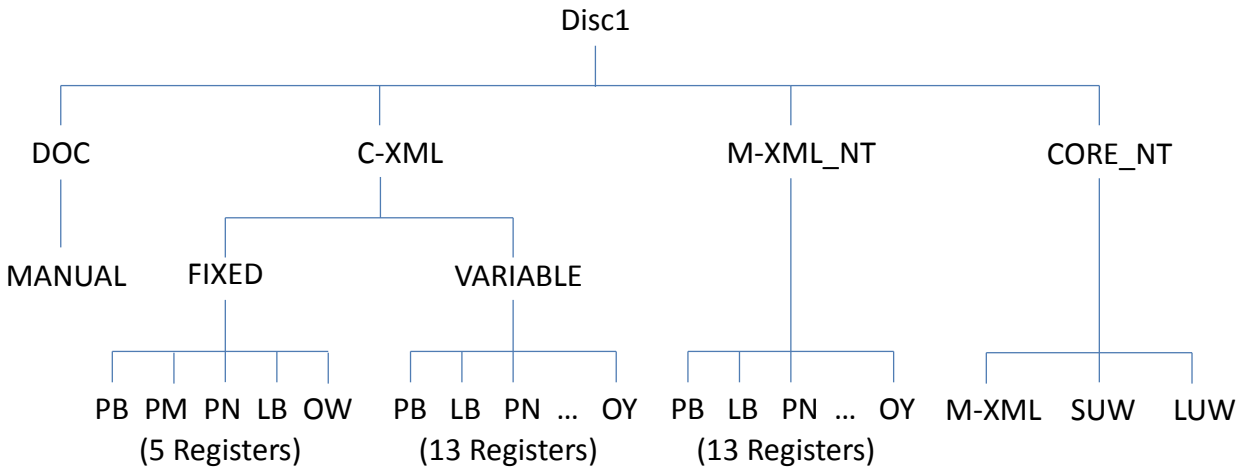

図 1-2A: BCCWJ-DVD 版（Version 1.1） Disc 1 のディレクトリ構成

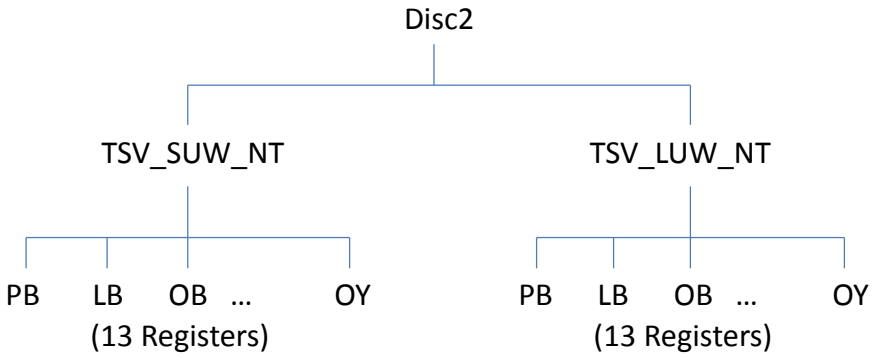

図 1-2B: BCCWJ-DVD （Version 1.1） Disc 2 のディレクトリ構成

Disc 3 と Disc 4 は、Version1.1 で新規に公開する非 NumTrans 版データを格納している。Disc 3（図 1-2C 参照）の M-XML_OT ディレクトリには M-XML の非 NumTrans 版が格納されており、CORE_OT ディレクトリには BCCWJ コアデータに含まれるサンプルの M-XML データと TSV データの非 NumTrans 版が格納されている[3]。前者は Disc 3 の M-XML_OT ディレクトリ内文書と、後者は後述する Disc 4 の TSV_SUW_OT、TSV_LUW_OT ディレクトリ内のデータと重複して格納されている。

最後に Disc 4 は、非 NumTrans 版の TSV データを保管している（図 1-2D 参照）。ディレクトリ構造は Disc 2 に準じている。

[3] ディレクトリ名に含まれる OT は original text の意味である。

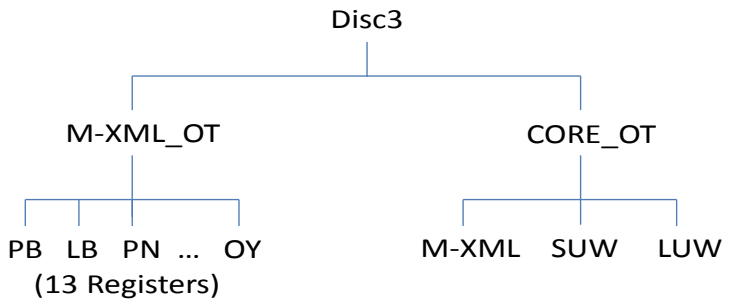

図 1-2C: BCCWJ-DVD 版（Version 1.1） Disc 3 のディレクトリ構成

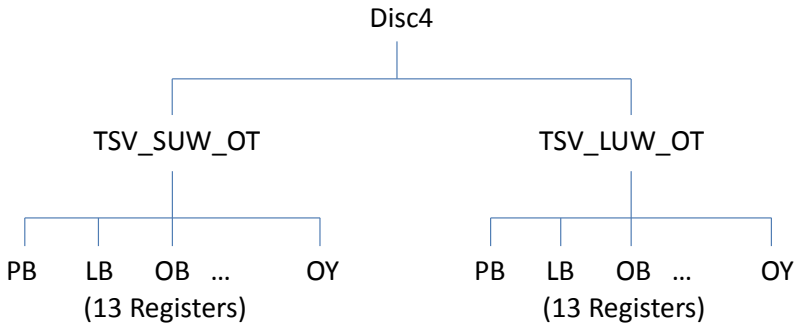

図 1-2D: BCCWJ-DVD 版（Version 1.1） Disc 4 のディレクトリ構成

これらのディスク中の圧縮ファイルを解凍すると、データサイズは数倍に増加するので、解凍時にはハードディスクに十分な残量を確保しておく必要がある。解凍前後でのデータサイズの変化を表 1-4A、B にまとめた。表 1-4A は XML 文書類の場合、表 1-4B は TSV データの場合をまとめており、表中の「前」「後」は「解凍前」「解凍後」の意味である。

PB（書籍）、LB（図書館 SC）、OC（Yahoo!知恵袋）、OY（Yahoo!ブログ）はファイル数、データ量が過大なので、圧縮に工夫を凝らしている。Disc 1 では、これらのディレクトリの圧縮ファイルを解凍すると複数のサブディレクトリに分けてファイルが格納される仕様になっている（表 1-4A、B でこれらのディレクトリの「後」はサブディレクトリ群を合計した値を示している）。

Disc 2、Disc 4 では、これらのディレクトリの圧縮ファイルを解凍すると TSV データが現れる。大部分のレジスターでは、そのレジスターの全データを含む 1 個のファイルが現れるだけであるが、LB と PB に関しては、TSV_SUW_NT、TSV_SUW_OT、TSV_LUW_NT、TSV_LUW_OT いずれも解凍後のデータサイズが 2GB を超えるので、ユーザーが利用している PC のファイルシステムが 2GB を超えるサイズのファイルに対応していない場合に配慮して、データを複数（5～20 個）のファイルに分割している。ユーザーはこれらのファイルを結合（concatenate）して当該レジスター用の TSV データを構成する必要がある。

表 1-4A: XML データのファイルサイズの解凍前後での変化（単位はメガバイト）

Register	C-XML				M-XML			
	FIXED		VARIABLE		NT		OT	
	前	後	前	後	前	後	前	後
PB*	20.5	59.2	63.6	243.0	1,157.6	9,597.0	1,153.3	9,583.4
PM	4.3	12.6	11.4	45.5	192.7	1517.3	191.6	1,516.8
PN	3.3	8.7	3.1	8.3	58.8	445.2	58.4	444.3
LB*	21.8	63.6	70.0	265.0	1,250.8	10,307.8	1,247.4	10,296.1
OB	--	--	9.3	37.1	155.4	1291.2	155.2	1,290.8
OW	2.9	8.0	8.9	35.4	181.3	1,513.7	178.8	1,503.9
OP	--	--	8.3	40.3	151.5	1,233.3	149.1	1,226.4
OL	--	--	1.4	7.8	34.2	322.0	34.2	322.0
OM	--	--	7.7	31.0	188.7	1,629.8	188.5	1,629.3
OT	--	--	2.3	9.2	37.4	317.4	37.1	316.5
OV	--	--	0.8	4.3	9.6	73.6	9.6	73.6
OC*	--	--	60.4	119.0	519.2	3,516.1	518.2	3,516.0
OY*	--	--	48.9	123.0	500.1	3,663.1	497.8	3,658.2
合計	52.8	152.1	296.2	968.9	4,437.0	35,427.6	4,419.1	35,377.2

*解凍後の値はサブディレクトないし複数ファイルにわけて格納されているデータの合計値

表 1-4B: TSV データのファイルサイズの解凍前後での変化（単位はメガバイト）

Register	NT				OT			
	SUW		LUW		SUW		LUW	
	前	後	前	後	前	後	前	後
PB*	864.4	4,823.8	617.6	3,572.9	842.1	4,827.4	591.6	3,568.2
PM	146.4	769.9	100.7	563.4	141.3	773.9	95.6	562.4
PN	44.0	229.1	29.5	161.2	43.0	229.5	28.4	160.9
LB*	930.7	5,112.0	672.7	3,862.3	911.4	5,113.2	647.6	3,858.5
OB	114.3	630.5	83.6	485.8	112.8	630.6	81.4	485.7
OW	139.2	820.4	91.4	535.1	132.0	820.9	85,2	532.1
OP	122.0	675.7	78.3	436.9	115.4	679.3	72.3	434.6
OL	24.4	173.6	16.1	114.7	24.3	173.6	16.1	114.7
OM	142.0	844.1	103.1	617.6	139.0	844.2	96.3	617.4
OT	27.9	160.1	20.2	118.9	26.9	160.2	19.2	118.7
OV	7.0	33.7	4.9	26.5	7.0	33.6	4.9	26.5
OC*	296.0	1,658.3	214.7	1,258.1	294.4	1,661.7	213.3	1,257.7
OY*	337.6	1,780.4	236.8	1,334.3	331.9	1,784.8	232.2	1,332.4
合計	3,195.9	17,711.7	2,269.7	13,087.8	3,121.6	17,732.9	2,183.9	1,3069.9

*解凍後の値はサブディレクトないし複数ファイルにわけて格納されているデータの合計値

1.4 BCCWJ-DVD 版の意義

『中納言』を利用できる環境にあるユーザーにとって、BCCWJ-DVD 版の存在意義はどこにあるだろうか。『中納言』は「語」（短単位ないし長単位）を単位としてコーパスを検索するツールである。語や語の連鎖を対象とした検索ならば、『中納言』でかなりのところまで用が足りる。

一方、『中納言』では検索できない情報もある。語の属性であっても現在の『中納言』では検索条件に指定できない属性が関与している場合（①,②）、「語」以外の単位が検索条件に関与している場合（③,④,⑤,⑥,⑦）、語ではなくサンプルの属性の検索（⑧,⑨）などは、『中納言』では実施不可能であるか、後処理を必要とする[4]。

① 特定の長さの語を検索する
② 和語だけを検索する
③ 文や段落の長さを測る
④ 文や段落の冒頭に生じやすい語を調査する
⑤ 個々のサンプルの語数を知る
⑥ サンプルごとに「ですます」体と「である」体の生起率を調べる
⑦ 常用漢字の出現頻度リストを作成する
⑧ 書き手の性別や年齢を検索条件に含めて語を検索する
⑨ 書き手の生年の分布を知る

BCCWJ-DVD 版を用いることによって、検索の可能性が大きくひらけてくる。ただしそれは検索に必要な情報を活用できるようになるという意味であって、万能の検索環境が提供されるという意味ではない。BCCWJ-DVD 版には検索ツール類は一切ふくまれていないので、ユーザーは自力で検索環境を構築する必要がある。本文書を読んで BCCWJ-DVD 版の購入を検討しているユーザーは、この点に特に留意していただきたい。

BCCWJ-DVD 版に適した検索環境は何かという問いあわせを受けることがある。ユーザーのスキルによって回答は異なってくるのだが、最も多くのユーザーに当てはまると考えられるのは、TSV 形式のデータはそのままの形でリレーショナルデータベース（RDB）にインポートできるので、MySQL、PostgreSQL、SQL Server などの RDB を利用して、SQL 言語で検索するのが便利ではないか、という回答であろう。

XML 文書を利用するためには、どうしてもある程度のプログラミングスキルが必要である。Ruby、Perl、Python 等のスクリプト言語でそれぞれの XML 処理用ライブラリを利用することが多いだろうが、XSLT のような XML 文書専用の言語もある。

---

[4] 後処理とは『中納言』の検索結果をダウンロードして、そこに含まれる情報を表計算ソフトやリレーショナルデータベース（RDB）などで集計する作業のことである。

1.5　BCCWJ の参考文献

BCCWJ は、構築途上で公開された数種類の「モニター版」も含めて、2011 年の公開以来、国内外の多くの研究者、研究機関によって利用されてきている。その結果、BCCWJ を参照・引用した研究文献も多数出版されている。本稿執筆の時点（2015 年 2 月）で確実に確認されているものだけで、内外 500 件以上の文献があり、国立国語研究所コーパス開発センターのホームページに文献リストが掲載されている[5]。

研究論文で BCCWJ を参照するにはどのような文献を引用すればよいかという問い合わせをもらうこともある。引用の目的によってどの文献が最適かは異なってくるが、以下にいくつか代表的な文献を紹介しておくことにする。

まず英文文献としては以下が代表的である。Disc 1 の MANUAL サブディレクトリにはこの論文の PDF が保管されている（LRE_2014.pdf）[6]。

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. "Balanced corpus of contemporary written Japanese". *Language Resources and Evaluation* 48 (2), pp.345-371 (DOI 10.1007/s10579-013-9261-0), 2014:06.

和文であれば、以下の書籍が代表的である。

山崎誠[編]『書き言葉コーパス ―設計と構築―』講座日本語コーパス 2, 朝倉書店, 2014 (ISBN978-4-254-51602-9 C3381).

この本は全 6 章と付録からなるが、そのうち以下の 5 章で BCCWJ の設計と構築に関する問題が多面的に論じられている。

第 1 章　コーパスの設計	[山崎誠・前川喜久雄]
第 2 章　サンプリング	[丸山岳彦・柏野和佳子]
第 3 章　文書構造の電子化	[山口昌也]
第 4 章　形態論情報	[小椋秀樹]
第 5 章　形態素解析	[小木曽智信]

本マニュアルを引用する場合は以下の書誌情報に準拠していただきたい。

---

[5] http://www.ninjal.ac.jp/corpus_center/bccwj/list.html　このリストは定期的にアップデートされる。

[6] この論文の扱いは Creative Commons Attribution 4.0 International (CC BY)に従う。

国立国語研究所コーパス開発センター「『現代日本語書き言葉均衡コーパス』利用の手引第 1.1 版」国立国語研究所, 2015.

Version 1.1 の Disc 1 の MANUAL サブディレクトリには、本マニュアルの他に BCCWJ の開発過程で蓄積された以下の作業用マニュアル類も保管されている[7]。BCCWJ の設計と構築の詳細情報はこれらの文献から得ることができる。

[1] 丸山岳彦・秋元祐哉「『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 －現代日本語書き言葉の文字数調査－」(JC-D-06-02.pdf)
[2] 丸山岳彦・秋元祐哉「『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法(2) －コーパスの設計とサンプルの無作為抽出法－」(JC-D-07-01.pdf)
[3] 柏野和佳子・丸山岳彦・稲益佐知子・田中弥生・秋元祐哉・佐野大樹・大矢内夢子・山崎誠「『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例」(JC-D-08-01.pdf)
[4] 丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子「『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用」(JC-D-01.pdf)
[5] 丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子「『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装」(JC-D-10-02.pdf)
[6] 高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也「JIS X 0213:2004 運用の検証」(JC-D-09-01.pdf)
[7] 西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也「『現代日本語書き言葉均衡コーパス』における電子化テキストの構築」(JC-D-10-03.pdf)
[8] 山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる「『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2」(JC-D-10-04.pdf)
[9] 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕「『現代日本語書き言葉均衡コーパス』形態論情報規程集 第４版 （上)」(JC-D-10-05-01.pdf)
[10] 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕「『現代日本語書き言葉均衡コーパス』形態論情報規程集 第４版 （下)」(JC-D-10-05-02.pdf)
[11] 小木曽智信・中村壮範「『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版」(JC-U-10-01.pdf)

[7] さらに多くの BCCWJ 関連文書が国立国語研究所コーパス開発センターのホームページで公開されている。
http://www.ninjal.ac.jp/corpus_center/bccwj/doc.html

1.6　BCCWJ 構築の経緯

1.6.1 Version 1.0 の公開まで

BCCWJ の構築は、その構想段階にまで遡ると 2004 年に始まった。同年春に『日本語話し言葉コーパス』の公開を終えた後、国立国語研究所研究開発部門（当時）の有志が集まって、コーパス利用の可能性を探るなかで、現代日本語を対象とした書き言葉均衡コーパスの必要性に対する認識が共有され、後に BCCWJ となる均衡コーパスの概念設計が始まった。翌 2005 年には文科省科学研究費（基盤研究 C, 課題番号 17632002, 代表者:前川喜久雄）の補助を得て、100 万語規模のパイロット版コーパスの構築実験を実施した。

BCCWJ の本格的な構築作業は、国立国語研究所のコーパス整備計画 KOTONOHA 計画の一部として 2006 年 4 月に 5 年計画で始まり、2011 年 7 月末に終了した。その間、2007 年末から 2009 年秋にかけては、独立行政法人に関する行政改革の一環として、国立国語研究所が独立行政法人から大学共同利用機関法人へと移管される騒動があり、BCCWJ 開発チームにもその影響が及んだ。しかし開発メンバーの結束と努力によって、オンライン版も DVD 版も大幅に遅延することなく公開を果たすことができたのは幸いであった。本章冒頭で述べたように Version 1.0 の DVD 版を公開したのは 2011 年 12 月のことであった。

BCCWJ の開発資金には、国立国語研究所の運営費交付金にくわえて、文科省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築:21 世紀の日本語研究の基盤整備」（略称、特定領域研究「日本語コーパス」、領域代表者：前川喜久雄、2006-2010 年度）の補助を受けた。両資金の分担関係としては、書籍に関するデータ（サンプル ID が PB、LB、OB で始まるサンプル、第 3 章参照）の構築に特定領域研究の研究費をあて、それ以外を運営費交付金でまかなった。

1.6.2 Version 1.1 における修正

BCCWJ-DVD 版（Version 1.0）の公開後、ユーザーから寄せられた意見のうち、早急な対応を必要としたのが、文境界の認定基準の見直しであった。書き言葉において文末を認定し、文境界を設定することは、句読点などの記号類が用いられている以上、容易であると思われるかもしれない。しかし、実際に 1 億語相当のサンプルを処理してみると、文末が記号類で明示されていないサンプルが頻出することにくわえ、複雑長大な引用の存在、果ては文末であるのか否かを文法的には解決不能と思われるサンプルの存在まで、複雑多岐な問題に直面する。Version 1.0 は 5 年間という強い時間的制約の下で開発したため、文境界認定の基準が十分に練り上げられておらず、問題の複雑さに対処しきれていなかった。

文境界認定の異同は、文数・文長などの計量言語学的指標に影響するだけでなく、係り受け構造や述語項構造などの言語アノテーション作業にも深刻な影響を及ぼす。そこで 2013 年初夏には国立国語研究所コーパス開発センターで、文境界認定基準の再検討を開始した。

その後、約 1 年の検討期間を経て、2014 年春には文境界修正方針の成案を得たので、実

際の修正作業に着手した。今回の修正で文末認定に関するすべての問題が解決されたわけではないが、Version 1.0 に比較すれば大幅に問題が軽減されているものと信じる。

また Version 1.1 の公開を機に非 NumTrans 版データも公開することにした。NumTrans は先に述べたように数字を形態素解析しやすくするための前処理であるが、短単位と数字の対応をとるためにすべての数字を漢字表記に変換する。もちろん原文の表記情報が失われているわけではなく、XML 文書中にタグを付して保存されているのだが、『中納言』その他でユーザーの目にとまるのが漢字に変換された文字列であるため、原文を改変してしまっているとの誤解を生じる原因となった。また自然言語処理の研究者からも、処理の煩雑さを厭う声があがっていた。非 NumTrans 版の公開によって、これらの批判にも前向きに応えることができたと信じる。

文境界認定基準の再検討には浅原正幸・小木曽智信・山口昌也・山崎誠・丸山岳彦・中村壮範・小西光・田中弥生と筆者が、その後のデータ修正作業には、上記にくわえて立花幸子・加藤祥・今田水穂・間淵洋子が参加した。

1.7 謝辞

サンプルの利用許諾をいただいた延べ 1 万人を超える個人著作権者のみなさまに、心より感謝申しあげる。

また先に 1.2.5 節で述べたように、BCCWJ の著作権処理では、多くの法人、団体のご協力をいただいた。以下にその名称を記して感謝のしるしとしたい。

公益社団法人日本文藝家協会、社団法人日本推理作家協会、社団法人日本児童文学者協会、社団法人日本児童文芸家協会、社団法人日本ペンクラブの各団体には、文芸分野でのサンプルの著作権者への広報および依頼状発送業務にご協力いただいた。また鷹羽狩行、篠弘の両氏には韻文関係のサンプル選定についてご指導をいただいた。

社団法人教科書協会、一般社団法人教学図書協会には、教科書出版各社との連絡を仲介していただいた。

一般社団法人日本音楽著作権協会には、歌詞に関係するサンプルの利用を許諾していただいた。

㈱朝日新聞社、㈱読売新聞社、㈱産業経済新聞社、㈱毎日新聞社、㈱京都新聞社、㈱中日新聞社、㈱高知新聞社、㈱神戸新聞社、㈱西日本新聞社、㈱北海道新聞社、㈱新潟日報社、㈱河北新報、㈱琉球新報社、㈱中国新聞社、一般社団法人共同通信社、㈱時事通信社からは新聞記事サンプルの利用を許諾していただいた。

ヤフー株式会社からは、Yahoo!知恵袋および Yahoo!ブログのデータを提供していただき、著作権の一括処理にご尽力いただいた。

白書の著作権処理に関しては中央省庁における担当部署に、また広報紙の著作権に関しては地方自治体の担当部署に、それぞれご協力いただいた。

衆議院記録部、参議院記録部、国会図書館の関係者からは国会会議録の著作権処理方針

について種々ご教示をいただいた。

個人著作権者との交渉に際しては、権利者との連絡をとるための窓口として、出版社に接触することが多かった。そのなかで、㈱アカデミー出版、㈱ヴィレッジブックス、㈱オライリー・ジャパン、㈱オレンジページ、㈱学習研究社、㈱経済界、㈱光人社、㈱小学館、㈱新潮社、㈱誠文堂新光社、㈱世界文化社、㈱ナツメ社、㈱南江堂、㈱日本実業出版社、㈱ハーレクイン、㈱PHP 研究所、㈱文芸社、㈱マガジンハウス、㈱みすず書房の各社においては格別に好意的なご対応をいただいた。

書籍、雑誌、新聞類の原本の閲覧、および書誌情報データの入手においては、大阪市立中央図書館、国立国会図書館、埼玉県立浦和図書館、埼玉県立久喜図書館、埼玉県立熊谷図書館、自治大学校図書室、湘北短期大学図書館、立川市図書館、東京都立多摩図書館、東京都立中央図書館、東京都立日比谷図書館、日本図書館協会、八王子市図書館、一橋大学附属図書館、横浜市中央図書館に便宜を図っていただいた。

付録：BCCWJ 開発メンバー

秋元祐哉	阿左美厚子	稲益佐知子	内元清貴	大石有香
大島一	大矢内夢子	小川志乃	小木曽智信	小椋秀樹
小沼悦	柏野和佳子	神野博子	河内昭浩	北村雅則
小磯花絵	小澤俊介	小西光	小林正行	小松祐美
近藤明日子	佐野大樹	鈴木翼	相馬さつき	高田智和
竹内ゆかり	田中牧郎	田中弥生	伝康晴	中村壮範
西部みちる	長谷川愛	服部龍太郎	原裕	平本智弥
平山允子	冨士池優美	前川喜久雄	間淵洋子	丸山岳彦
宮内佐夜香	舞木右	森本祥子	山口昌也	山崎誠
山田篤	吉田谷幸宏	渡部涼子	浅原正幸†	今田水穂†
加藤祥†	立花幸子†			

†Version 1.1 から参加